

Heterogeneous information network model for equipment-standard system

Liang Yin^a, Li-Chen Shi^a, Jun-Yan Zhao^a, Song-Yang Du^a, Wen-Bo Xie^{c,d},
Duan-Bing Chen^{b,c,d,*}

^a*Beijing Special Vehicle Institute, Beijing 100072, People's Republic of China*

^b*The Center for Digitized Culture and Media, UESTC, Chengdu 611731, People's
Republic of China*

^c*Big Data Research Center, University of Electronic Science and Technology of China,
Chengdu 611731, People's Republic of China*

^d*Web Sciences Center, University of Electronic Science and Technology of China,
Chengdu 611731, People's Republic of China.*

Abstract

Entity information network is used to describe structural relationships between entities. Taking advantage of its extension and heterogeneity, entity information network is more and more widely applied to relationship modeling. Recent years, lots of researches about entity information network modeling have been proposed, while seldom of them concentrate on equipment-standard system with properties of multi-layer, multi-dimension and multi-scale. In order to efficiently deal with some complex issues in equipment-standard system such as standard revising, standard controlling, and production designing, a heterogeneous information network model for equipment-standard system is proposed in this paper. Three types of entities and six types of relationships are considered in the proposed model. Correspondingly, several different similarity-measuring methods are used in the modeling process. The experiments show that the heterogeneous information network model established in this paper can reflect relationships between entities accurately. Meanwhile, the modeling process has a good performance on time consumption.

*Corresponding author at: Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China.

Email address: dbchen@uestc.edu.cn (Duan-Bing Chen)

Keywords: Complex system, Heterogeneous information network,
Equipment-standard system, Entity relationships model

1. Introduction

Complex network theory has been proven to be a powerful framework to understand the structure and dynamics of complex systems[1, 2, 3, 4, 5, 6, 7]. Entity information network is a kind of complex network that describes the structural relationships between entities. With more and more researches about entity information network model having been proposed, it is widely used in social network analyzing[8, 9, 10, 11], image association and other fields [12, 13]. Depending on the diversity of the relationships and entities, entity information network model can be divided into two categories: based on simple structure and based on complex one.

Generally, only one kind of entity relationship is contained in simple structure based model, such as item-to-item, object-to-item or object-to-object relationships.

In recent years, many modeling methods on item-to-item information network are proposed. In refs [14, 15], the similarity between media sources is evaluated by mapping different types of medium’s features to a common space based on media contextual clues. Zhu et al. [16] proposed an information network model to correlate tweets, emotion features and users based on emotion analysis.

Matrix transformation, matrix decomposition and random walk methods are used to construct object-to-item information network [17, 18, 19]. And some researches focus on the status of users in the resource, such as influence, importance and opinion leaders, by constructing the authoritative network based on specific topics [20]. Some modeling methods are also widely concentrated on object-to-object relationships such as user-to-user. Based on direct and indirect users’ relationships, the similarity between users can be measured by the comments on social media resource and those on user reviews [21, 22]. Besides, matrix decomposition can also be used to evaluate the similarity between users in social media combining with LDA [23].

Complex structure based model supports multiple forms of relationships between multimodal entities. Google’s Knowledge Graph [24] and other engine knowledge maps belong to this type of model. In which, there are multi-class of relationships between entities. In addition, the entities are

also multi-dimension and multi-scale. Thus, they are generally called heterogeneous information network models [25].

Although heterogeneous information network models are widely studied in complex systems, such as academic resource search [26], citations recommendation [27], user based personalized service [28, 29], traveling plan search and recommendation [30], and makeup recommendation [31]. There are seldom researches about equipment-standard system.

As we know, equipment-standard system is a multi-layer, multi-dimension and multi-scale complex system. For this reason, we present a heterogeneous information network model for equipment-standard system (HINM-ESS) in this paper. HINM-ESS contains three types of nodes that present different granularity of entities in equipment-standard system and six types of entity relationships. A complete HINM-ESS can provide strong support for equipment-standard system, such as resource searching, production designing, standard revising and controlling.

Two real data sets are used in experiments to verify the validity of HINM-ESS. The one is a real equipment-standard system data set that contains 2600 standard documents and 24 elements. The other is a mixed test data set that contains different size of data from multiple fields. The experiments show that our methods in modeling process are efficient and accurate. Comparing with Word Mover’s Distance (WMD) [32], the relational modeling between documents using our method can save 50% time and the performance of precision reduces about 20%. That is, we can establish HINM-ESS efficiently, and reflect relationship between entities in equipment-standard system accurately.

2. Model and Method

2.1. Framework of HINM-ESS

The formal expression of HINM-ESS is described as $HINM-ESS=(V, E)$. In which, $V = \{Doc, Item, Topic\}$ is the network node set with three different granularity of entities, i.e., *Doc*, *Item* and *Topic*. *Doc* represents the standard document; *Item* represents the clause in standard document; *Topic* represents the unit such as equipment, module or element. $E = \{E_{DD}, E_{DI}, E_{DT}, E_{II}, E_{IT}, E_{TT}\}$ is the network edge set with six different kinds of relationships between entities, where E_{DD} represents the *Doc* – *Doc* relationships, E_{DI} represents the *Doc* – *Item* relationships, E_{DT} represents

the *Doc* – *Topic* relationships, E_{II} represents the *Item* – *Item* relationships, E_{IT} represents the *Item* – *Topic* relationships, and E_{TT} represents the *Topic* – *Topic* relationships. Each edge has its weight to measure the degree of correlation or the similarity between entities.

As shown in Fig. 1, to construct the HINM-ESS, six kinds of entity relationships are confirmed in turn by evaluating the similarities or correlations between entities. First of all, weights of E_{DD} are evaluated to confirm the *Doc* – *Doc* relationships, as shown in Fig. 1(a). Secondly, a *Doc* is divided into several items to confirm *Doc* – *Item* relationships E_{DI} , as shown in Fig. 1(b). And then *Item* – *Item* relationships E_{II} and *Topic* – *Topic* relationships E_{TT} are confirmed by the same strategy used in the first step, as shown in Fig. 1(c). Thirdly, the *Item* – *Topic* relationships E_{IT} can be confirmed since *Item* – *Item* and *Topic* – *Topic* relationships have been confirmed, as shown in Fig. 1(d), and *Doc* – *Topic* relationships E_{DT} can be confirmed in virtue of E_{DD} and E_{TT} , as shown in Fig. 1(e). Finally, HINM-ESS can be obtained, as shown in Fig. 1(f). Because these six relationships involve different entities, the method to measure the weights of edges are very different. The details will be described in following subsections.

2.2. *Doc* – *Doc* relational modeling

The *Doc* – *Doc* relational modeling is the first and the most important step to establish the HINM-ESS. As the contents of *Docs* are text, the correlation between *Docs* can be confirmed via text similarity so as to establish the connection between two *Doc* nodes in HINM-ESS. Since the contents of *Items* and *Topics* are text as well, only *Doc* – *Doc* relational modeling is described in details, the *Item* – *Item* and *Topic* – *Topic* relational modeling is similar with *Doc* – *Doc*.

2.2.1. *Docs* similarity calculation via WMD

Text analyzing is one of the most popular research topics. There are many researches focus on measuring text similarity, such as LDA [33] and word2vec [34]. These models translate text contents to different abstract features to improve the measuring performance. In this paper, we choose Word Mover’s Distance (WMD) [32] method to measure the similarity between *Docs*.

WMD provides accurate similarity measurement by combining the word embedding with the Earth Mover’s Distance (EMD). In WMD, word2vec provides embedding matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ for a finite size vocabulary of n words. The i^{th} column, $\vec{x}_i \in \mathbf{R}^d$, represents the embedding of the i^{th} word in d –

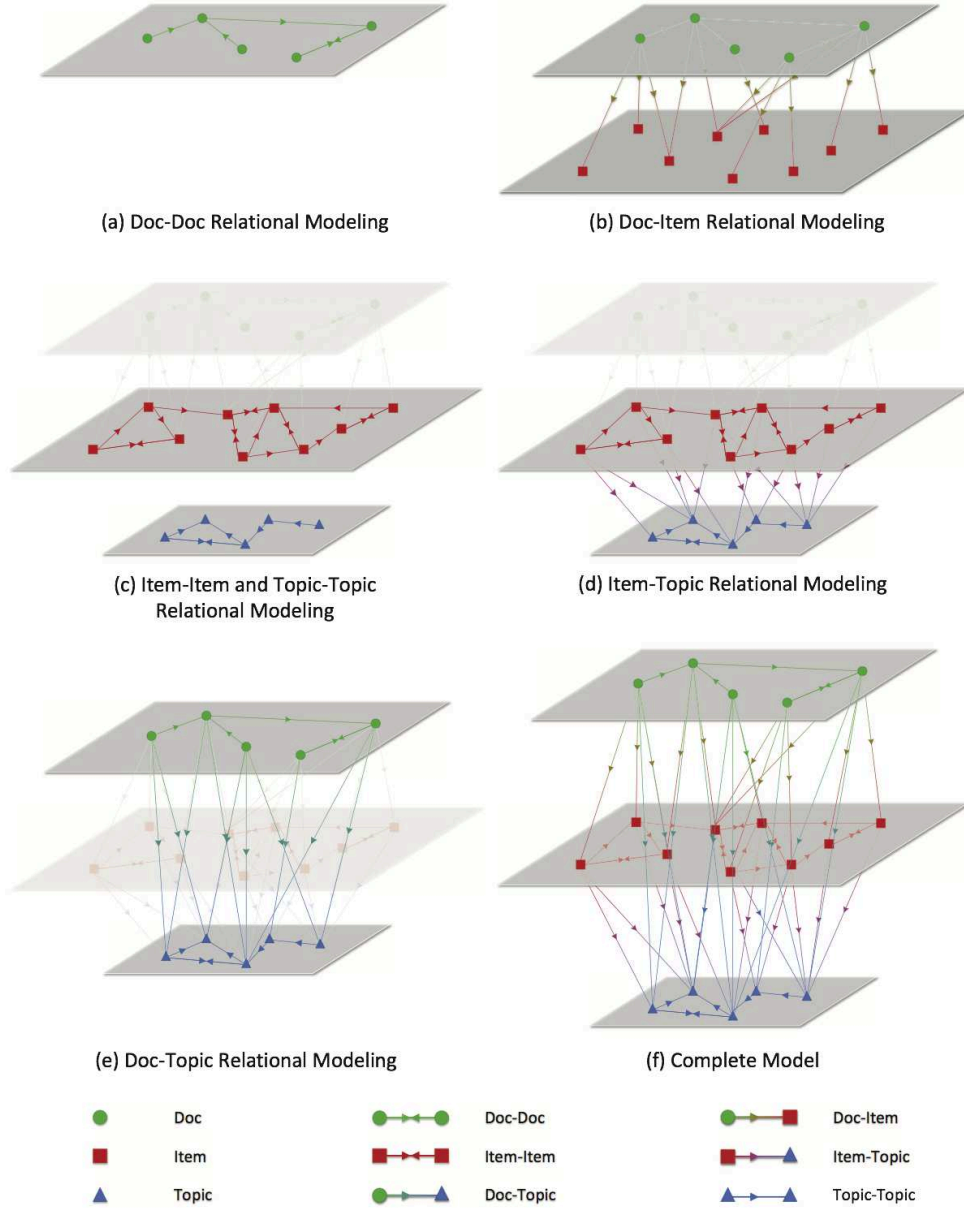


Figure 1: The framework of HINM-ESS modeling.

dimensional space. Therefore, semantic similarity between two words w_i and w_j , that is refer to as the cost associated with ‘traveling’ from one word to another, is measured by their Euclidean distance in the word2vec embedding space, that is,

$$cost_{ij} = \|\vec{x}_i - \vec{x}_j\|_2. \quad (1)$$

Text documents are represented as normalized bag-of-words (nBOW) vectors, $\vec{d} \in \mathbf{R}^n$, if word w_i appears c_i times in the document, the i^{th} element d_i in \vec{d} is defined as,

$$d_i = \frac{c_i}{\sum_{j=1}^n c_j}. \quad (2)$$

Let \vec{d} and \vec{d}' be the nBOW representation of two text documents, $T \in \mathbb{R}^{n \times n}$ be a flow matrix denotes ‘how much’ of word w_i in \vec{d} travels to word w_j in \vec{d}' , The similarity evaluation problem can be defined as the minimum cumulative cost required to move all words from \vec{d} to \vec{d}' , and the constraints are provided by the solution to the following linear program,

$$\begin{aligned} & \min_{T_{ij} \geq 0} \sum_{i,j=1}^n T_{ij} \cdot cost_{ij}, \\ & \text{subject to} \quad \sum_{j=1}^n T_{ij} = d_i \quad \forall i \in \{1, 2, \dots, n\}, \\ & \quad \quad \quad \sum_{i=1}^n T_{ij} = d'_j \quad \forall j \in \{1, 2, \dots, n\}. \end{aligned} \quad (3)$$

After the similarities between *Docs* being calculated by WMD, we regard these similarity values as the weights of E_{DD} .

Although WMD leads to low error rates, the time complexity is as high as $O(n^3 \log n)$. In order to improve the time consumption while maintaining the accuracy, we modify WMD by introducing SimHash [35] strategy. In which, Top- N potentially similar *Docs* are screened via SimHash to reduce the complexity significantly.

2.2.2. Top- N potentially similar *Docs* screening

The main idea of SimHash strategy is to reduce the dimension. As a local sensitive Hash method, SimHash maps the high-dimensional eigenvectors to a signature value containing f bits named f -bit fingerprint. By comparing

the Hamming distance between f -bit fingerprints, we can estimate whether the *Docs* are similar or not. Therefore, an efficient potentially similar *Docs* screening strategy based on SimHash is presented in this paper.

In Top- N potentially similar *Docs* screening process, SimHash is used to map f -bit fingerprints for each *Doc*. The i^{th} *Doc* D_i in dataset is mapped

to f -bit fingerprints with number 0 or 1, for instance, $\overbrace{(0101\dots00111)}^{f \text{ bits}}$. Based on the fingerprints, the Hamming distance is chosen to estimate the similarity between *Docs* and generate a list of Top- N potentially similar *Docs* efficiently. As shown in Eq. 4, Hamming distance is the number of '1' in XOR between documents D_n and D_m .

$$H_{nm} = D_n \oplus D_m. \quad (4)$$

For instance, the Hamming distance between 110 and 011 equals 2 since $110 \oplus 011 = 101$.

After estimating the similarity between *Docs* by Hamming distance, we can obtain the Top- N potentially similar documents list by directly using sorting algorithms. But this simple strategy is very time-consuming. Even for the best sorting algorithm, its time complexity reaches $O(n \log n)$. Aiming at the Top- N *Docs*, there is no need to sort all Hamming distances. Therefore, we propose two strategies to improve the efficiency of potentially similar document lists generating process.

(1) The lowliest replace elimination based strategy.

In this strategy, Top- N potentially similar documents are stored in a finite set of k elements, which is defined as $SET = \{D_i | 0 \leq i < k\}$. The update strategy of SET is

$$SET = \begin{cases} (SET \cup D_\Delta) \setminus \{\max(SET)\}, & Ham(\max(SET)) > Ham(D_\Delta) \\ SET, & Ham(\max(SET)) \leq Ham(D_\Delta) \end{cases} \quad (5)$$

where $\max(SET)$ represents the node which is the farthest from the target node on hamming distance in SET ; D_Δ represents a new node which is under judgment to be the potentially similar *Docs*; $Ham(\cdot)$ represents the hamming distance between this node and target node.

(2) Ordered window filling based strategy.

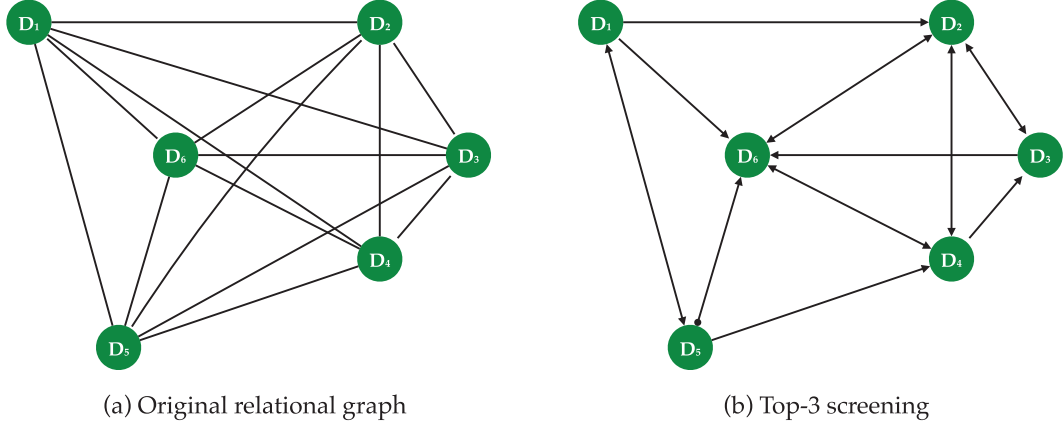


Figure 2: *Doc - Doc* relational modeling processes.

The lowest replace elimination based strategy needs to repeatedly update the maximum value in the finite set. It takes too much time to traverse the whole set. Therefore, we propose another strategy based on ordered window filling to reduce traversal time further. In this strategy, Top- N potentially similar documents are stored in an ordered window of k elements, which is defined as $WIN = \{D_i | 0 \leq i < k, Ham(D_i) \leq Ham(D_{i+1})\}$. So, the elements D_i in WIN can be updated via Eq. 6 under the condition of $Ham(D_m) \leq Ham(D_\Delta) \leq Ham(D_{m+1})$,

$$D_i = \begin{cases} D_{i-1}, & m+1 < i \leq k \\ D_\Delta, & i = m+1 \\ D_i, & 0 \leq i < m \end{cases} \quad (6)$$

in which, D_Δ represent a new node which is under judgment to be the potentially similar *Docs*; k is the size of WIN .

Via the screen strategies above, we obtain the Top- N potentially similar *Docs* by removing weightless edges from original *Doc - Doc* relational graph. This process is illustrated by the simple example in Fig. 2.

Firstly, we have a lot of edges in the original *Doc - Doc* relational graph that is a complete graph, as shown in Fig. 2(a). By screening the Top-3 potentially similar *Docs*, we remove the weightless edges in Fig. 2(a), and then obtain the *Doc - Doc* relational graph, as shown in Fig. 2(b). The edges are directed since two *Docs* may not be the Top- N similar for each other at the same time.

As there are only a small number of edges which link the potentially similar documents left, the time consumption of the similarity evaluation process with WMD method is reduced sharply. Then *Doc – Doc* relational graph can be completed efficiently via two stages, i.e., Top-*N* potentially similar *Doc* screening and Docs’ similarity evaluation via WMD method.

2.3. *Doc – Item Relational modeling*

Each *Doc* contains different numbers of *Items*. Therefore, the *Doc – Item* relational modeling can be considered as a decomposition process. The core issue of the *Doc – Item* relational modeling process is to extract items from standard documents accurately.

Since *Docs* in equipment-standard system have typical hierarchical structure, the section numbers in section headers are always composed of integers and symbolic points, ‘2.2’ for instance. Furthermore, the section number will be followed by the section title directly. Therefore, all the possible section numbers and titles can be extracted via regular expressions to construct a triplet sequence $T_i = (Cap_i, No_i, In_i)$. In which, *Cap* is the chapter number that is the first integer of section number (for example, ‘2’ is the chapter number when the section number is ‘2.2’); *No* is the whole section number; *In* is the line number of title in the document. For instance, a possible triplet may like this: (5, 5.2.1, 456).

The results obtained via regular expressions still contain some noises that meet the definition of the section title but not the real one, such as data in the table or in the text of the reference data. For this reason, we design following noise-filtering rule to remove those illogical triplets,

$$\begin{cases} No_{i-1} < No_i \\ \max \{Cap_j | j < i\} \leq Cap_i, \\ In_{i-1} < In_i \end{cases} \quad (7)$$

in which, section number must conform the typesetting format of sections (when $No_{i-1} = 2.2$, the logical value of No_i can only be 2.2.1 or 2.3 or 3); the chapter number Cap_i must be no less than all the chapter numbers in the previous triplets. In_i in the triplet sequence must be sorted.

According to the noise-filtering rule, most of noises are erased. As a result, we can use the line number in triplets to decompose the *Docs* and to extract out the *Items* accurately. At the same time, the *Doc – Item* relational network is constructed.

79 standard documents have been used to test the *Item* extracting process. There are 77 documents extracted correct completely. The other two documents have only 1 incorrect item respectively. The accuracy of the *Items* extracting process is 97.5%.

2.4. *Item – Topic relational modeling*

In HINM-ESS, two relationships, *Doc – Topic* and *Item – Topic*, are widely used in equipment-standard system application. Just *Item – Topic* relational modeling process is analyzed in this paper, since *Doc – Topic* relational modeling is similar.

Generally, there are part of *Items* have been assigned *Topic* labels, that is, there are some known *Item – Topic* relationships in equipment-standard system. Hence, these relationships can be used to measure the correlation between *Items* and *Topics* indirectly, and then complete the *Item – Topic* relational model.

For a new *Item* I_i that is waiting for assigning *Topic* label, the correlation W_{I_i, T_k} between I_i and the k^{th} *Topic* T_k is measured by

$$W_{I_i, T_k} = \frac{\sum_{I_j \in S_{I_i}} W_{I_i, I_j} \cdot W_{I_j, T_k}}{|S_{I_i}|}, \quad (8)$$

where I_j is one of I_i 's similar *Items*; W_{I_i, I_j} is the similarity between I_i and its similar *Item* I_j which obtained in *Item – Item* relational modeling; S_{I_i} presents I_i 's similar *Items* set; W_{I_j, T_k} presents the correlations between I_j 's similar *Items* and their relative *Topics*.

Analogously, for a new *Topic* T_k , the correlation between T_k and the i^{th} *Item* I_i is measured by combining with T_k 's similar *Topics* and their relative *Items*,

$$W_{I_i, T_k} = \frac{\sum_{T_p \in S_{T_k}} W_{I_i, T_p} \cdot W_{T_p, T_k}}{|S_{T_k}|}, \quad (9)$$

where T_p is one of T_k 's similar *Topics*; W_{T_p, T_k} is the similarity between T_k and its similar *Topic* T_p ; S_{T_k} presents T_p 's similar *Topics* set; W_{I_i, T_p} is the correlations between T_k 's similar *Topics* and their relative *Items*.

The *Item – Topic* relational modeling process, which is the core of the HINM-ESS, is illustrated in Fig. 3. Figure 3(a) shows four *Items*, five *Topics* and five known relationships. In original graph, I_3 and T_5 are waiting

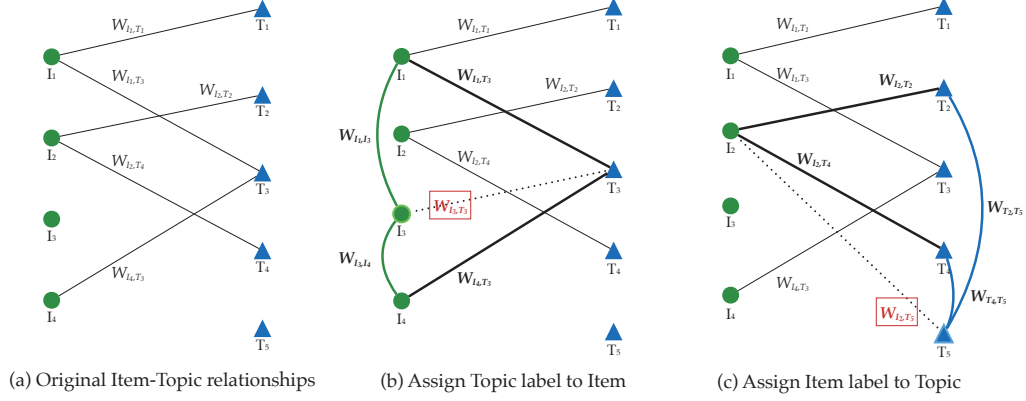


Figure 3: Examples of *Item – Topic* relational modeling.

for assignment. Since I_1 and I_4 are assigned to T_3 , the correlation between I_3 and T_3 is measured by

$$W_{I_3, T_3} = \frac{W_{I_3, I_1} \cdot W_{I_1, T_3} + W_{I_3, I_4} \cdot W_{I_4, T_3}}{2}, \quad (10)$$

as shown in Fig. 3(b). Similarly, the correlations W_{I_2, T_2} and W_{I_2, T_4} are known, therefore, the correlation between I_2 and T_5 is measured by,

$$W_{I_2, T_5} = \frac{W_{I_2, T_2} \cdot W_{T_2, T_5} + W_{I_2, T_4} \cdot W_{T_4, T_5}}{2}, \quad (11)$$

as shown in Fig. 3(c).

After all the correlations are evaluated, the weightless *Item – Topic* edges will be removed according to a threshold θ . Alike soft-classification method, multiple *Topics* will be assigned to one *Item* in *Item – Topic* relational modeling process and vice versa.

In the *Item – Topic* relational model, if the relationship between a *Item* and *Topic* is confirmed, they will be put into the training samples. This iterative process will optimize HINM-ESS continually and make *Items – Topic* model more and more accurate.

3. Results and Discussions

In this paper, the performance of *Doc – Doc* and *Item – Topic* relational network modeling methods are tested on two real data (DATA1 and DATA2).

DATA1 includes 2600 standard documents and 24 elements. DATA2 is a mixed-field text data set which contains different size of data (from 1KB to 1GB).

The performance of original WMD and SimHash+WMD is compared on DATA1 through three indices. The first index is Time Improvement (TI)

$$TI = \frac{T_{WMD} - T_{SimHash+WMD}}{T_{WMD}}, \quad (12)$$

where T_{WMD} represents the time consumption of WMD algorithm and $T_{SimHash+WMD}$ represents the time consumption of SimHash+WMD method.

The second index is accuracy that is used to estimate the proportion of documents SimHash mistakenly delete from the 20 most similar text evaluated by WMD,

$$Accuracy_N = \frac{|\min_{20}(WMD) \cap \min_{20}(SimHash + WMD)|}{20}, \quad (13)$$

where $Accuracy_N$ is the accuracy when SimHash method screening top- N documents as potentially similar documents; $\min_{20}(WMD)$ is the set that contains the 20 most similar documents evaluated by WMD and $\min_{20}(SimHash + WMD)$ is that by SimHash+WMD.

The third index is $F_1 - score$ that considers the time improvement and accuracy at the same time. It is defined as

$$F_1 - score = \frac{2 \times TI \times Accuracy}{TI + Accuracy}. \quad (14)$$

Figure 4 shows that TI decreases with Top- N increasing while the accuracy increasing with Top- N increasing. Since the $F_1 - score$ achieves maximum at the Top-1500, it is recommended to choose Top-1500 to screen potential similar *Docs* for 50% time saving and 20% precision reducing.

We also compare the time consumption of two improved screen strategies (*lowliest replace elimination based strategy* and *ordered window filling based strategy*) on DATA2. As shown in Fig. 5, the *lowliest replace elimination based strategy* only save 1% time while the *ordered window filling based strategy* can save about 7.5% running time in screening process. The $E - Time$ index in Fig. 5 is a time index enhanced by the time consumption of all hamming distances sort strategy,

$$E - time = \frac{T}{T_0}, \quad (15)$$

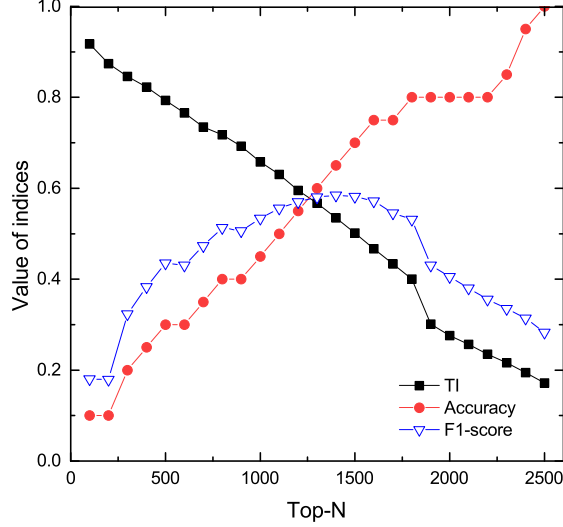


Figure 4: Time consumption compare on WMD and SimHash+WMD.

where T is the time consumption of *lowliest replace elimination based strategy* or *ordered window filling based strategy*; T_0 is that of whole sort strategy.

Finally, we test the HINM-ESS’s accuracy on relationship between entities using DATA1. We test on 8 *Topics* to verify whether the *Items* are linked to the right *Topics* or not. The precision equals to the ratio of the number of test *Items* and the number of correct *Items*. Generally, more training samples lead to higher precision. The precision on Topic #4 and #6 is higher than that on other Topics, since the number of training *Items* in these two *Topics* is about third times larger than that in others, as shown in Table 1. In addition, the size of *Items*’ training sample sets are larger than that of *Docs*’. Hence, as anticipated, the performance of precision on *Item – Topic* reflecting is much higher than that of *Docs*.

4. Conclusions

Suffering from the complex and redundant equipment-standard system, a heterogeneous information network model for equipment-standard system(HINM-ESS) is presented in this paper to deal with some important issues in the

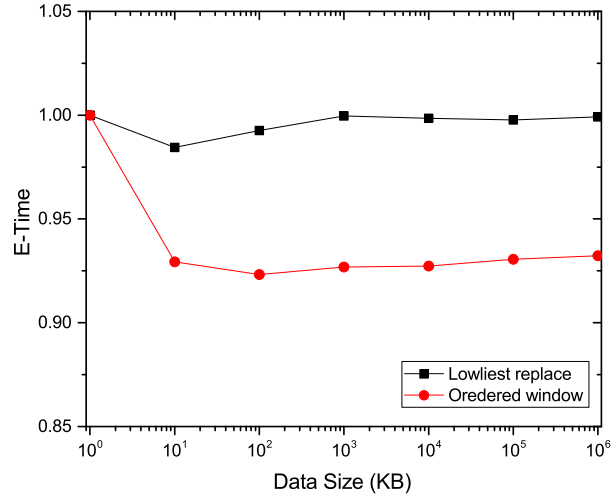


Figure 5: $E - Time$ on two different strategies in WMD+SimHash.

Table 1: Accuracy of *Item - Topic* relational model on DATA1.

Topic	#Training Item	#Test Item	#Correct Item	Precision(%)
#1	143	17	15	88.24
#2	209	25	24	96.00
#3	132	16	13	81.25
#4	585	67	67	100
#5	116	14	13	92.86
#6	530	60	60	100
#7	62	8	7	87.50
#8	90	11	11	100
Average	233	27	26	96.30

system, such as standard documents searching, standard revising, standard controlling and, production designing. HINM-ESS contains three types of nodes that represent three types of entities and six types of entity relationships. *Doc – Doc*, *Doc – Item*, *Item – Item*, *Doc – Topic*, *Item – Topic*, and *Topic – Topic* relational models are discussed and the detail modeling strategies are presented in this paper.

Experiments on two real data sets show that the modeling strategies are time saving and the accuracy is also rather good. Moreover, experiments show that HINM-ESS model can reflect real *Item – Topic* relationships accurately when training sample is enough. Overall, HINM-ESS is an efficient and accurate model. It will provide a strong and firm support for applications in equipment-standard system.

We will consider the iterative strategy in the modeling process to improve the accuracy of relational models further in the future study.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China under Grant Nos. 61433014 and 61673085, and by the Fundamental Research Funds for the Central Universities under Grant No. ZYGX2014Z002.

References

References

- [1] R. Albert, A.L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* 74 (2002) 47-97.
- [2] S.N. Dorogovtsevyz, J.F.F. Mendes, Evolution of networks, *Adv. Phys.* 51 (2002) 1079-1187.
- [3] M.E.J. Newman, The structure and function of complex networks, *SIAM Rev.* 45 (2003) 167-256.
- [4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.U. Hwang, Complex networks: Structure and dynamics, *Phys. Rep.* 424 (2006) 175-308.
- [5] L.D.F. Costa, F.A. Rodrigues, G. Travieso, P.R. Villas Boas, Characterization of complex networks: A survey of measurements, *Adv. Phys.* 56 (2007) 167-242.

- [6] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, Critical phenomena in complex networks. *Rev. Mod. Phys.* 80 (2008) 1275-1335.
- [7] M. Barthélemy, Spatial networks, *Phys. Rep.* 499 (2011) 1-101.
- [8] A.X. Cui, Z.K. Zhang, M. Tang, P.M. Hui, Y. Fu, Emergence of scale-free close-knit friendship structure in online social networks, *PLoS ONE* 7 (2012) e50702.
- [9] L. Lü, Y.-C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case, *PLoS ONE* 6 (2011) e21202.
- [10] T. Zhou, M. Medo, G. Cimini, Z.-K. Zhang, Y.-C. Zhang, Emergence of scale-free leadership structure in social recommender systems, *PLoS ONE* 6 (2011) e20648.
- [11] S.E. Ahnert, T.M.A. Fink, Clustering signatures classify directed networks, *Phys. Rev. E* 78 (2008) 036112.
- [12] L. Lü, T. Zhou, Link prediction in complex networks: A survey, *Physica A* 390 (2011) 1150-1170.
- [13] S. Maslov, K. Sneppen, A. Zaliznyak, Detection of topological patterns in complex networks: correlation profile of the internet, *Physica A* 333 (2004) 529-540.
- [14] H. Zhang, J. Yuan, X. Gao, Z. Chen, Boosting cross-media retrieval via visual-auditory feature analysis and relevance feedback, In: *Proceedings of the 22nd ACM International Conference on Multimedia*, ACM, 2014, pp. 953-956.
- [15] Y. Gao, M. Wang, Z.J. Zha, J. Shen, X. Li, X. Wu, Visual-textual joint relevance learning for tag-based social image search, *IEEE T. Image Process.* 22 (2013) 363-76.
- [16] L. Zhu, A. Galstyan, J. Cheng, K. Lerman, Tripartite graph clustering for dynamic sentiment analysis on social media, In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ACM, 2014, pp. 1531-1542.

- [17] M. Clements, A.P.D. Vries, M.J.T. Reinders, The influence of personalization on tag query length in social media search, *Inform. Process. Manag.* 46 (2010) 403-412.
- [18] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, J. Han, Personalized entity recommendation: a heterogeneous information network approach, In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, ACM, 2014, pp. 283-292.
- [19] J. Vosecky, K.W.T. Leung, W. Ng, Collaborative personalized Twitter search with topic-language models, In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, 2014, pp. 53-62.
- [20] Y. Li, C. Wu, X. Wang, P. Luo, A network-based and multi-parameter model for finding influential authors, *Journal of Informetrics* 8 (2014) 791-799.
- [21] A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec, Effects of user similarity in social media, In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ACM, 2012, pp. 703-712.
- [22] H. Ma, On measuring social friend interest similarities in recommender systems, In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, 2014, pp. 465-474.
- [23] Z. Wang, J. Liao, Q. Cao, H. Qi, Z. Wang, Friendbook: A semantic-based friend recommendation system for social networks, *IEEE T. Mobile Comput.* 14 (2015) 538-551.
- [24] K.J. Vang, Ethics of Google's Knowledge Graph: some considerations, *J. Inf. Commun. Ethics Soc.* 11 (2013) 245-260.
- [25] Y. Sun, J. Han, *Meta-Path-Based Similarity Search and Mining, Mining Heterogeneous Information Networks: Principles and Methodologies*, Morhan & Claypool, 2012, pp.55-94.

- [26] M.-F. Chiang, J.-J. Liou, J.-L. Wang, W.-C. Peng, M.-K. Shan, Exploring heterogeneous information networks and random walk with restart for academic search, *Knowl. Inf. Syst.* 36 (2013) 59-82.
- [27] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, J. Han, Clus-cite: Effective citation recommendation by information network-based clustering, In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2014, pp. 821-830.
- [28] A. Kao, W. Ferng, S. Poteet, L. Quach, R. Tjoelker, TALISON - Tensor analysis of social media data, *IEEE International Conference on Intelligence and Security Informatics*, IEEE, 2013, pp. 137-142.
- [29] C. Li, A. Sun, Fine-grained location extraction from tweets with temporal awareness, In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, 2014, pp. 43-52.
- [30] A. J. Cheng, Y. Y. Chen, Y.T. Huang, W.H. Hsu, H.Y.M. Liao, Personalized travel recommendation by mining people attributes from community-contributed photos, In: *Proceedings of the 19th ACM International Conference on Multimedia*, ACM, 2011, pp. 83-92.
- [31] L. Liu, J. Xing, S. Liu, H. Xu, X. Zhou, S. Yan, Wow! you are so beautiful today!, *ACM T. Multim. Comput.* 11 (2014) 20.
- [32] M. J. Kusner, Y. Sun, N.I. Kolkin, K. Q. Weinberger, From word embeddings to document distances, In: *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 957-966.
- [33] T.L. Griffiths, M. Steyvers, Finding scientific topics, *Proc. Natl. Acad. Sci.* 101 (2004) 5228-5235.
- [34] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, *arXiv 1301.3781* (2013).
- [35] C. Sadowski, G. Levin, Simhash: Hash-based similarity detection, <http://www.googlecode.com/sun/trunk/paper/SimHashwithBib.pdf>, 2007.